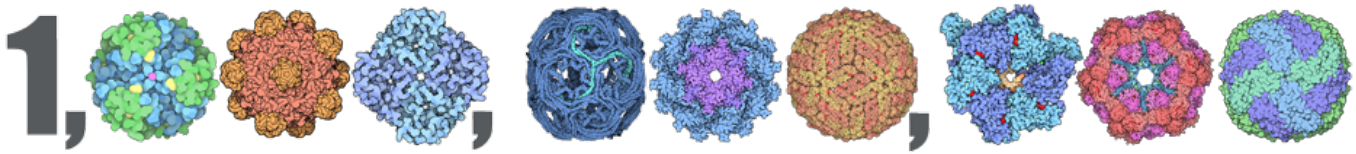
Introduction to MMTF-Spark

Peter W. Rose, Anthony R. Bradley, Jose M. Duarte,
Antonin Pavelka, Andreas Prlić,
Alexander S. Rose, Yana Valasatava

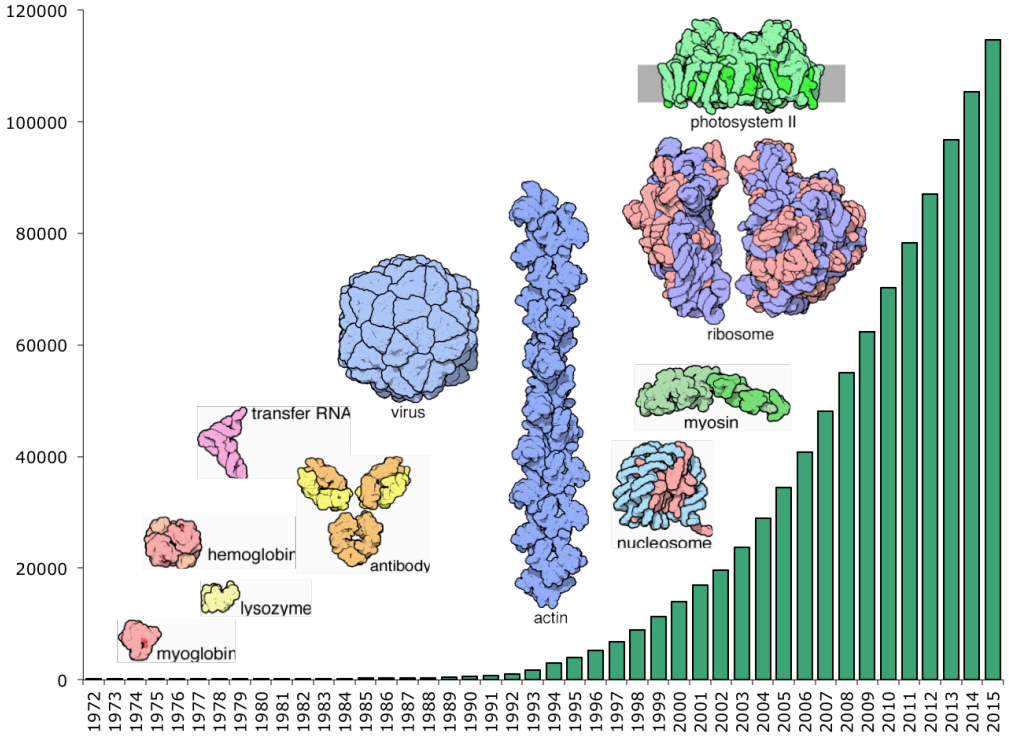
Structural Bioinformatics Laboratory
San Diego Supercomputer Center
UC San Diego

SDSC SAN DIEGO SUPERCOMPUTER CENTER RCSB PDB UC San Diego

PDB – A Billion Atom Archive

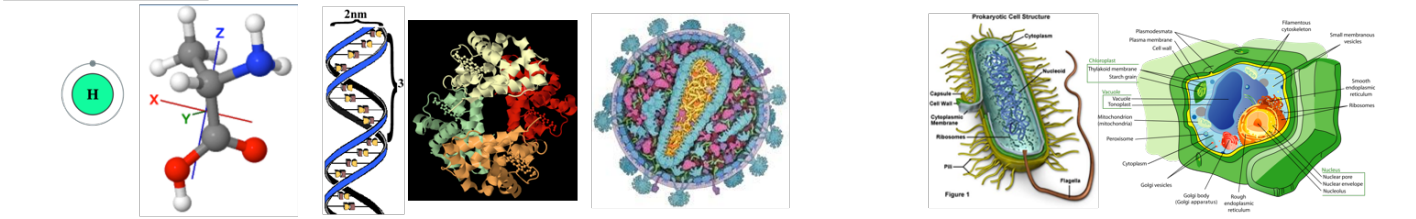


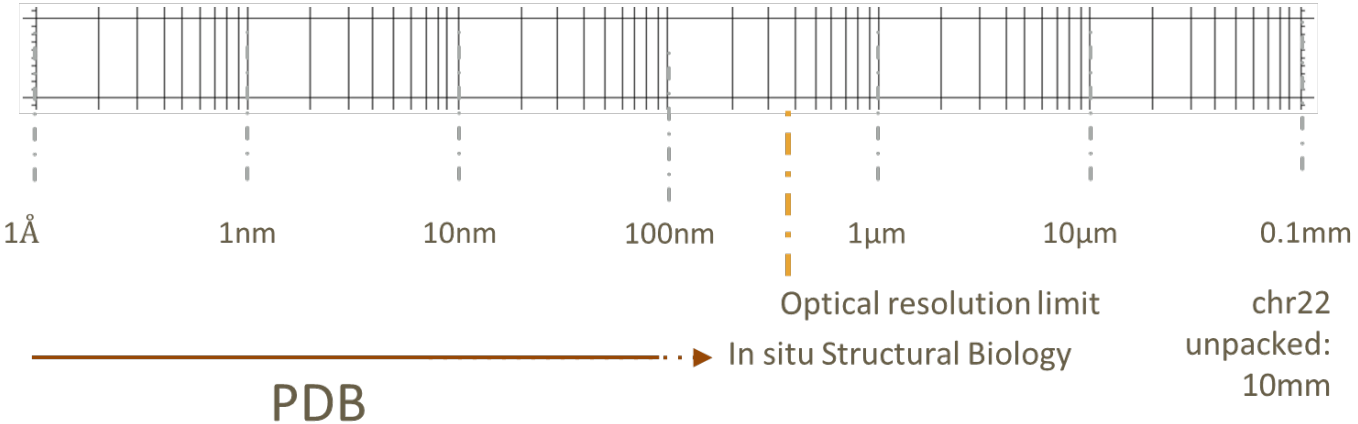
> 1 billion atoms in the asymmetric units



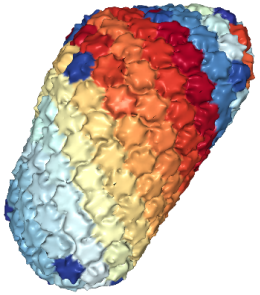
120,000 structures in June 2016

Growing Structure Size and Complexity



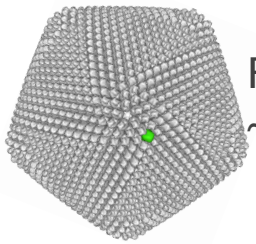


Largest asymmetric structure in PDB



HIV-1 capsid: PDB ID 3J3Q
~2.4M unique atoms

Largest symmetric structure in PDB



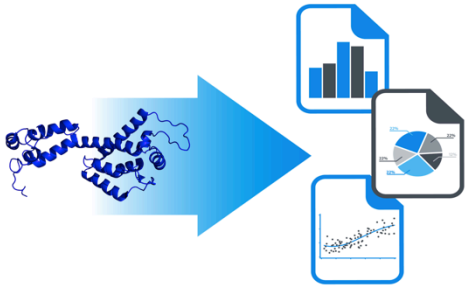
Faustovirus major capsid: PDB ID 5J7V
~40M overall atoms

→ Scalability Issues

- **Interactive visualization**
  - slow network transfer
  - slow parsing
  - slow rendering

- **Mobile visualization**
  - limited bandwidth
  - limited memory

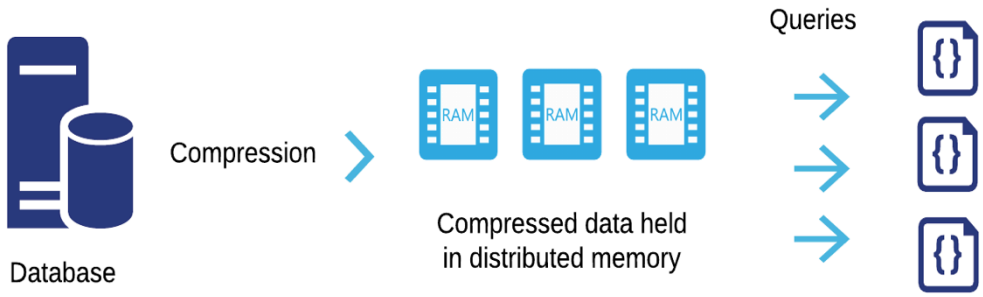
- **Large-scale structural analysis**
  - slow repeated I/O
  - slow repeated parsing



Compressive Structural Bioinformatics

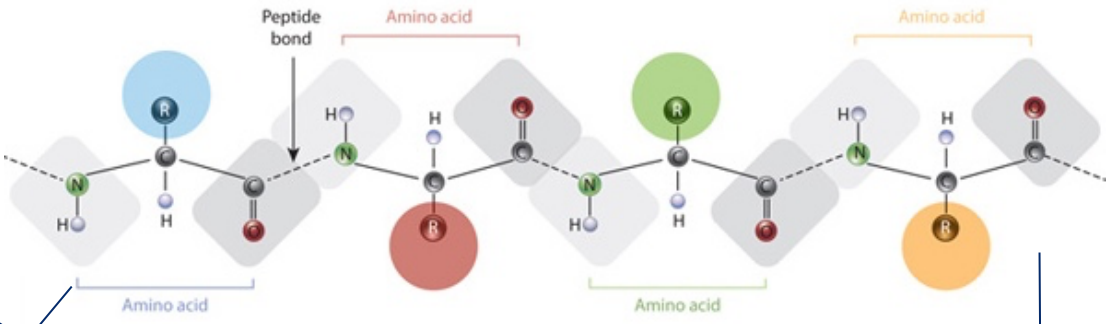
Efficiently store, transmit, and visualize 3D structures of biological macromolecules

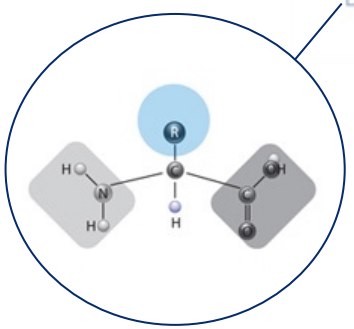
Perform large-scale structural calculations such as geometric queries or structural comparisons over the entire PDB archive held in memory

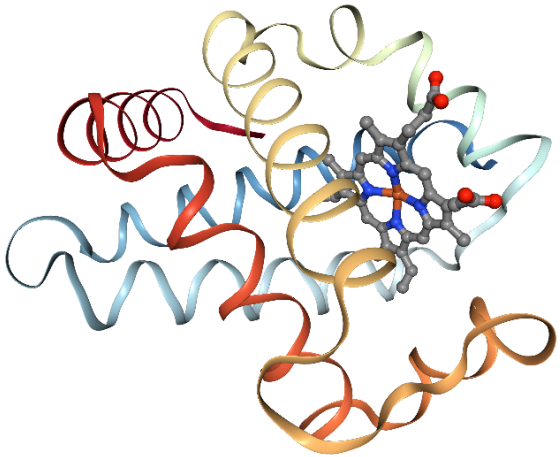


Macromolecular 3D Structure

Biological macromolecules: proteins, nucleic acids







Biological macromolecules are polymers constructed by linking monomers by covalent bonds

PDBx/mmCIF

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.Cartn_x_esd
_atom_site.Cartn_y_esd
_atom_site.Cartn_z_esd
_atom_site.occupancy_esd
_atom_site.B_iso_or_equiv_esd
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1    N  N   . TRP A 1 5   ? 8.519   -0.751  10.738  1.00 13.37 ? ? ? ? ? ? 5   TRP A N   1
ATOM   2    C  CA  . TRP A 1 5   ? 7.743   -1.668  11.585  1.00 13.42 ? ? ? ? ? ? 5   TRP A CA  1
ATOM   3    C  C   . TRP A 1 5   ? 6.786   -2.502  10.667  1.00 13.47 ? ? ? ? ? ? 5   TRP A C   1
ATOM   4    O  O   . TRP A 1 5   ? 6.422   -2.085  9.607   1.00 13.57 ? ? ? ? ? ? 5   TRP A O   1
ATOM   5    C  CB  . TRP A 1 5   ? 6.997   -0.917  12.645  1.00 13.34 ? ? ? ? ? ? 5   TRP A CB  1
ATOM   6    C  CG  . TRP A 1 5   ? 5.784   -0.209  12.221  1.00 13.40 ? ? ? ? ? ? 5   TRP A CG  1
ATOM   7    C  CD1 . TRP A 1 5   ? 5.681   1.084   11.797  1.00 13.29 ? ? ? ? ? ? 5   TRP A CD1 1
ATOM   8    C  CD2 . TRP A 1 5   ? 4.417   -0.667  12.221  1.00 13.34 ? ? ? ? ? ? 5   TRP A CD2 1
ATOM   9    N  NE1 . TRP A 1 5   ? 4.388   1.418   11.515  1.00 13.30 ? ? ? ? ? ? 5   TRP A NE1 1
ATOM   10   C  CE2 . TRP A 1 5   ? 3.588   0.375   11.797  1.00 13.35 ? ? ? ? ? ? 5   TRP A CE2 1
ATOM   11   C  CE3 . TRP A 1 5   ? 3.837   -1.877  12.645  1.00 13.39 ? ? ? ? ? ? 5   TRP A CE3 1
ATOM   12   C  CZ2 . TRP A 1 5   ? 2.216   0.208   11.656  1.00 13.39 ? ? ? ? ? ? 5   TRP A CZ2 1
ATOM   13   C  CZ3 . TRP A 1 5   ? 2.465   -2.043  12.504  1.00 13.33 ? ? ? ? ? ? 5   TRP A CZ3 1
ATOM   14   C  CH2 . TRP A 1 5   ? 1.654   -1.001  12.009  1.00 13.34 ? ? ? ? ? ? 5   TRP A CH2 1
```

Flexible, extensible, and verbose format with rich metadata, well suited for archival purposes (mmcif.wwpdb.org)

redundant annotations

inefficient representation

repetitive information

- **MacroMolecular Transmission Format (mmtf.rcsb.org)**
  - Compact
    - fast network transfer, less I/O
  - Fast to parse
    - binary, no string parsing
  - Contains information for structural analysis and visualization
    - covalent bonds and bond orders
    - consistently calculated secondary structure

Data In MMTF File

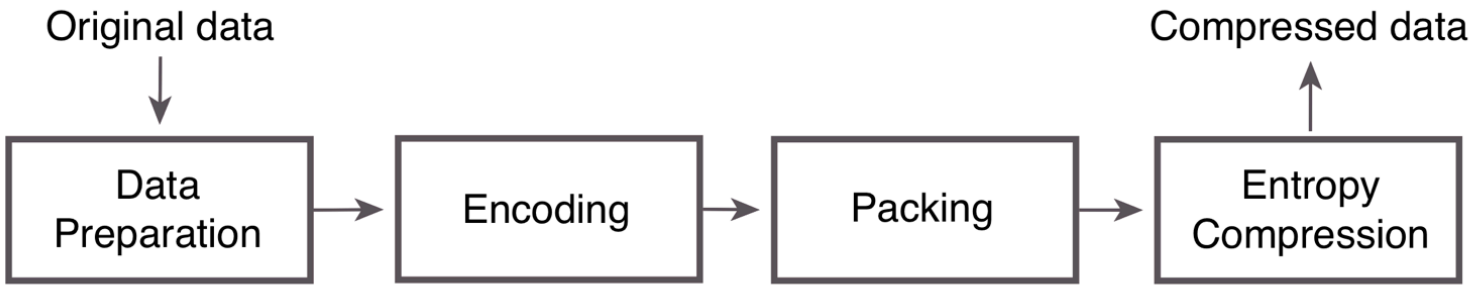
MMTF contains a subset of data used by visualization and structural analysis programs

Data Category	Data Items
Metadata	PDB ID, title, deposition date, release date, experimental method(s)
Crystallographic info	Space group, unit cell, NCS operators, resolution, R_{free}, R_{work}
Primary structure	Polymer sequences
Secondary structure	DSSP secondary structure assignments*
Structural model	Models, chains, groups (residues), atoms, bonds* and bond orders*
Quaternary structure	Biological assembly transformations

* These data items are not available in the PDBx/mmCIF files and are added to MMTF files.

https://doi.org/10.1371/journal.pcbi.1005575.t001

MMTF Compression Pipeline



extract structural data
calculate bonds, SSE

integer encoding
dictionary encoding
run-length encoding
delta encoding

recursive indexing

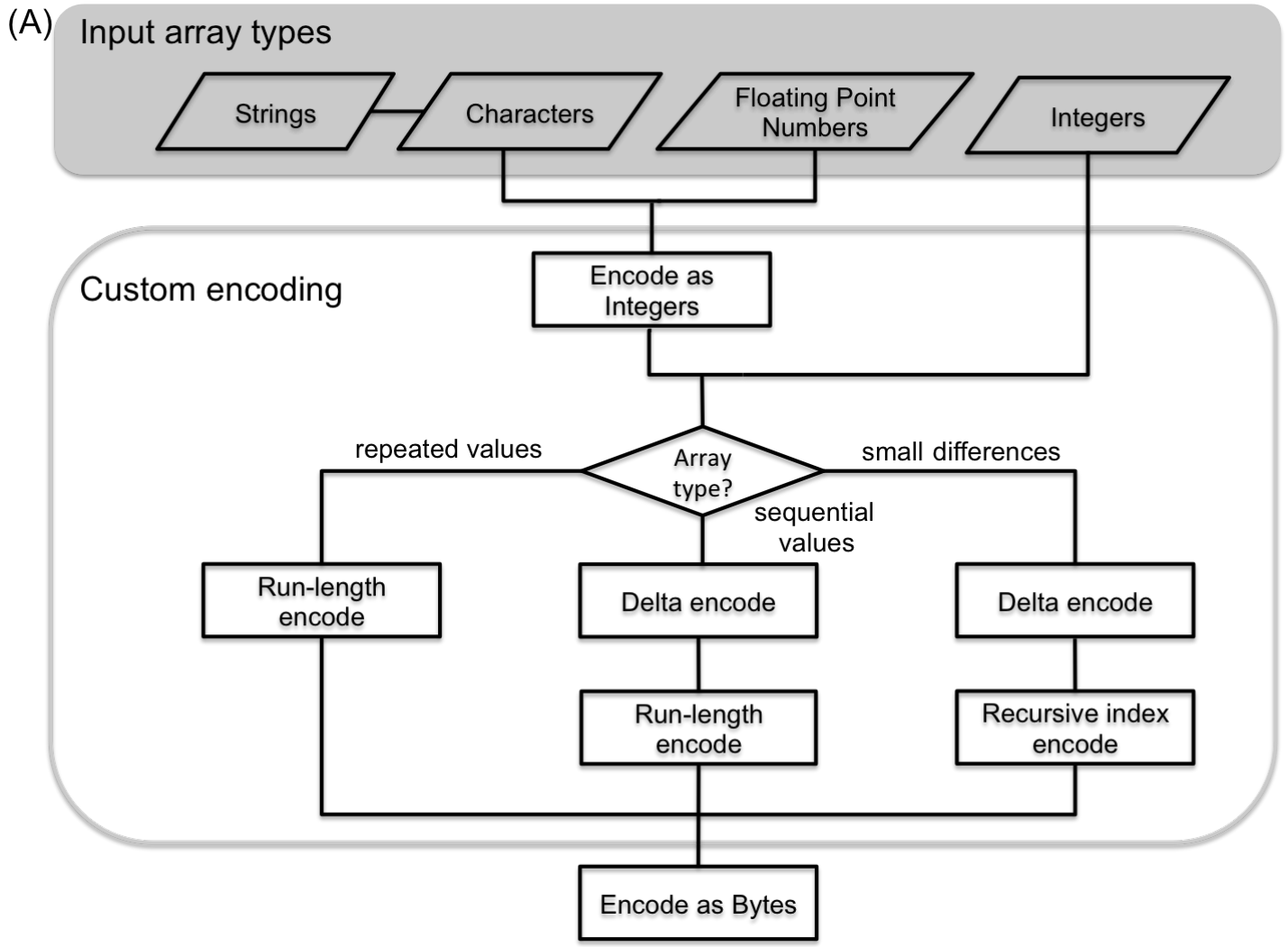
GZIP

Binary, extensible container format of MMTF

MessagePack

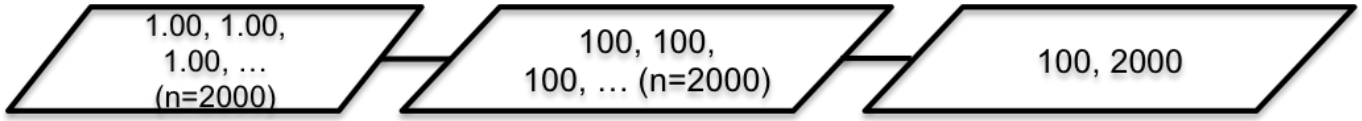
It's like JSON.
but fast and small.

Columnar Encoding of Arrays

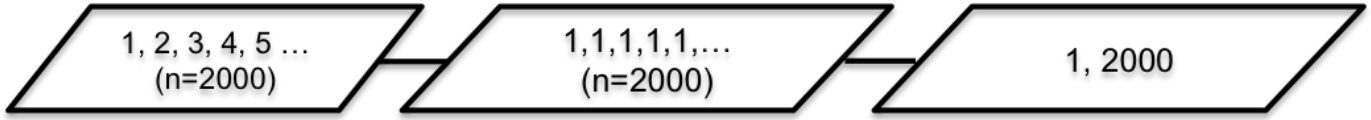


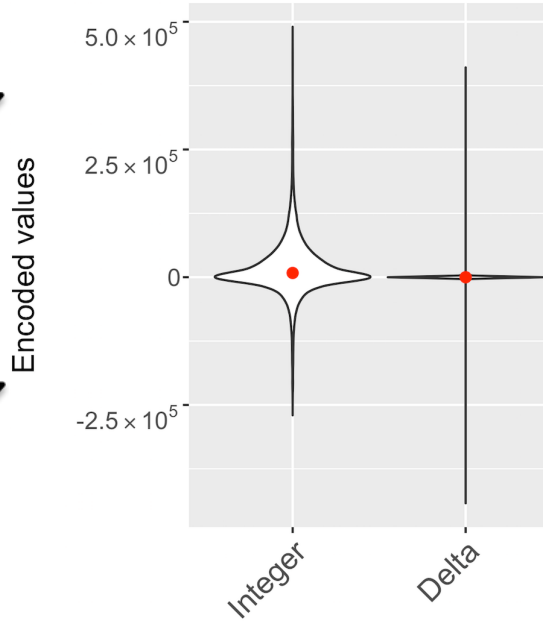
Custom Encoding Examples

(B) Repeated values — *Run-length encoding (e.g., occupancy)*



(C) Sequential values — *Delta and run-length encoding (e.g., serial numbers)*





(D) Small differences — *Integer, delta, and recursive index encoding (e.g., xyz-coordinates as 16-bit integers)*

Delta encoding reduced the dynamic range of numbers which makes them more compressible

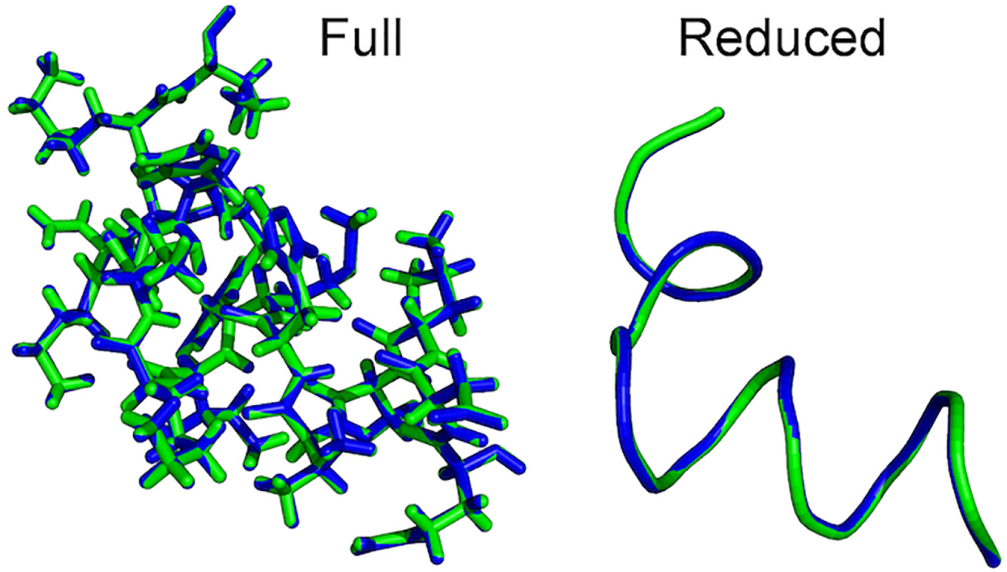
Dictionary Encoding

Unique groups (residues) are stored in a “dictionary”


```
{
        "groupName": "SER",
        "singleLetterCode": "S",
        "chemCompType": "L-PEPTIDE LINKING",
        "atomNameList": [ "N", "CA", "C", "O", "CB", "OG" ],
        "elementList": [ "N", "C", "C", "O", "C", "O" ],
        "formalChargeList": [ 0, 0, 0, 0, 0, 0 ],
        "bondAtomList": [ 1, 0, 2, 1, 3, 2, 4, 1, 5, 4 ],
        "bondOrderList": [ 1, 1, 2, 1, 1 ]
}
```


Unique entities are stored in a “dictionary”
(e.g., polymer type, polymer sequences)

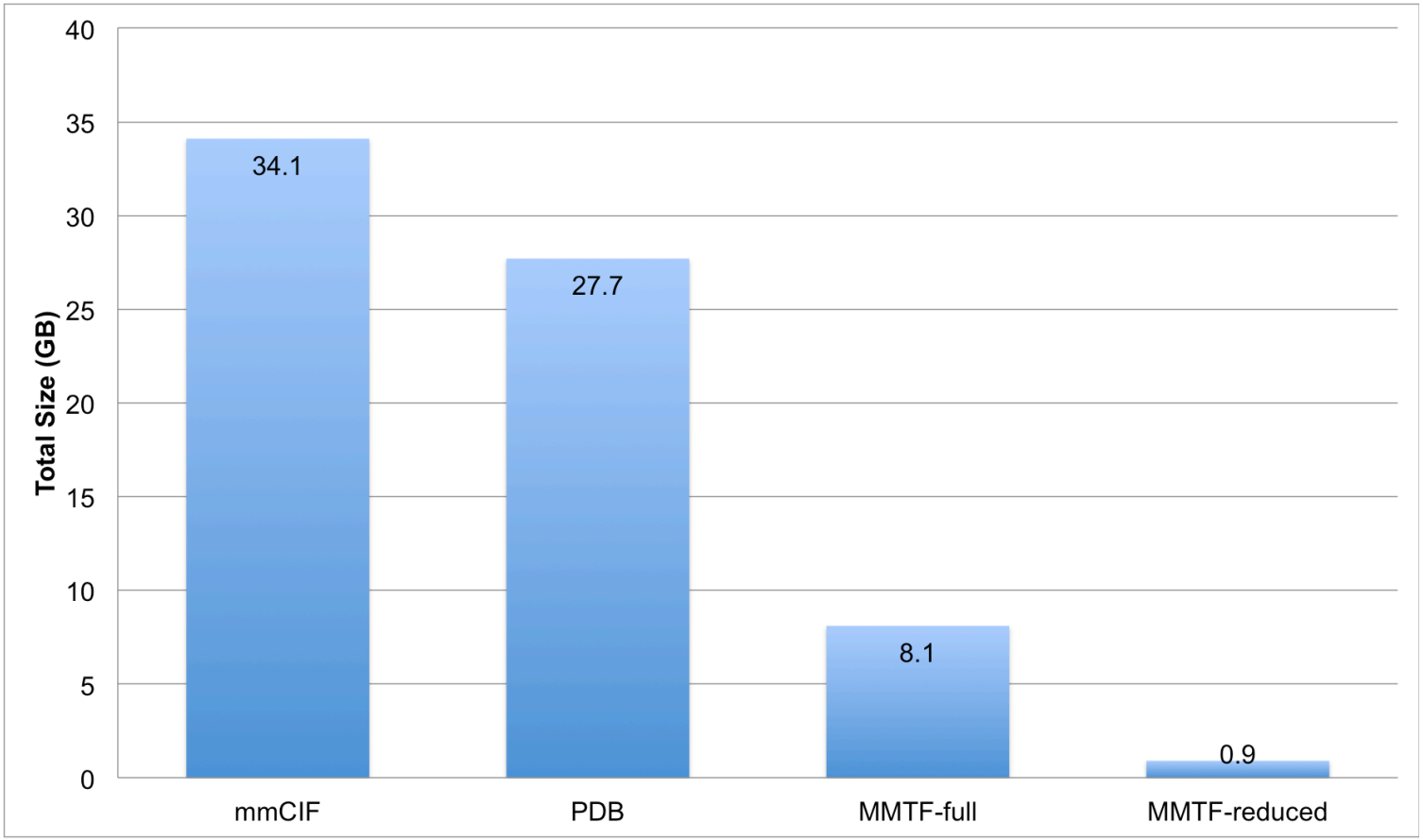
Lossless vs. Lossy Compression



Type	Representation	Coordinate precision (Å)	Temperature factor (Å^2) and Occupancy precision
full	all atoms	0.001	0.01
reduced	C-alpha (polypeptide), P (polynucleotide), all atoms (non-polymers)	0.1	0.1

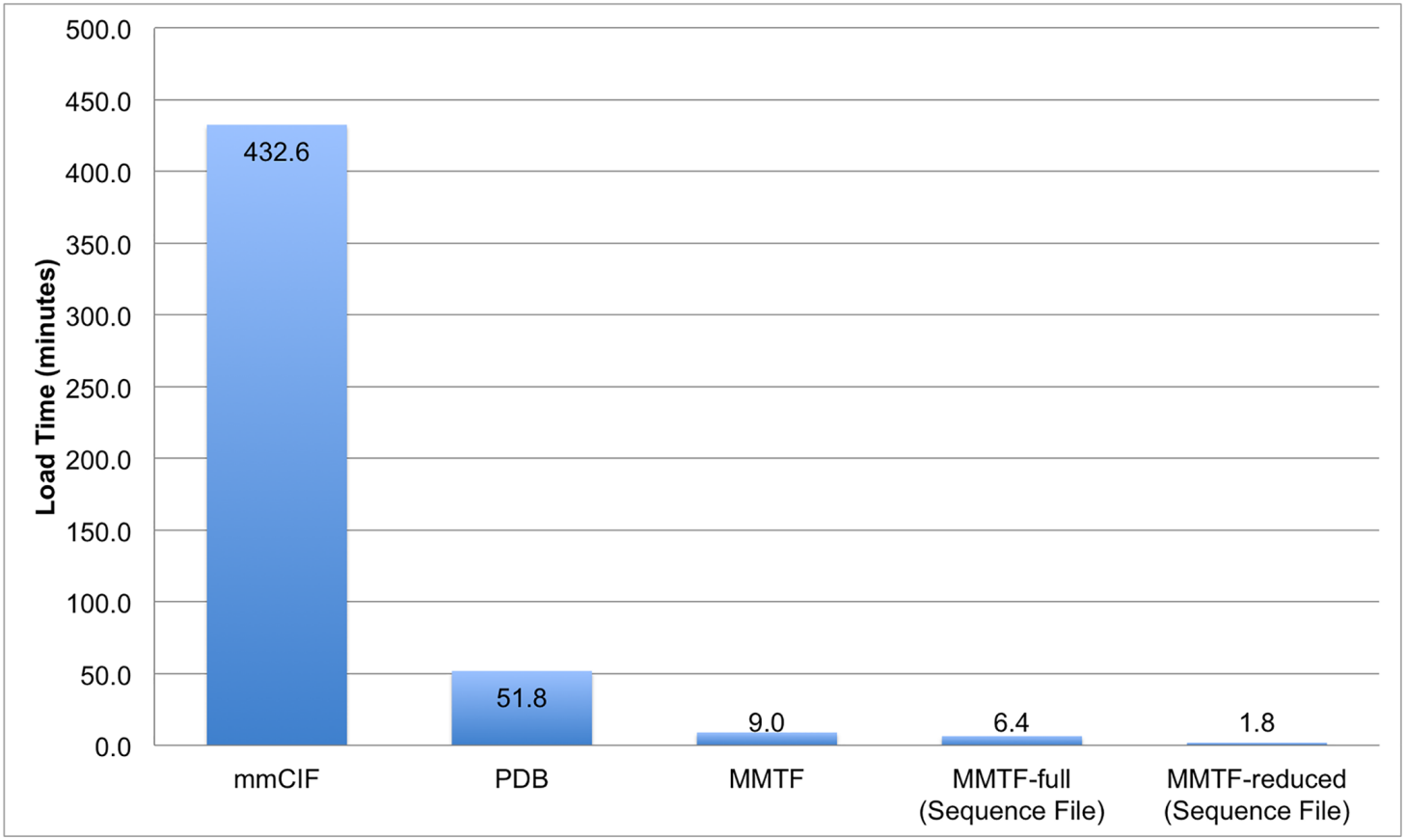
https://doi.org/10.1371/journal.pcbi.1005575.t002

File Size Comparison (PDB archive ~127,000 entries)



All file sizes are for the gzip compressed files

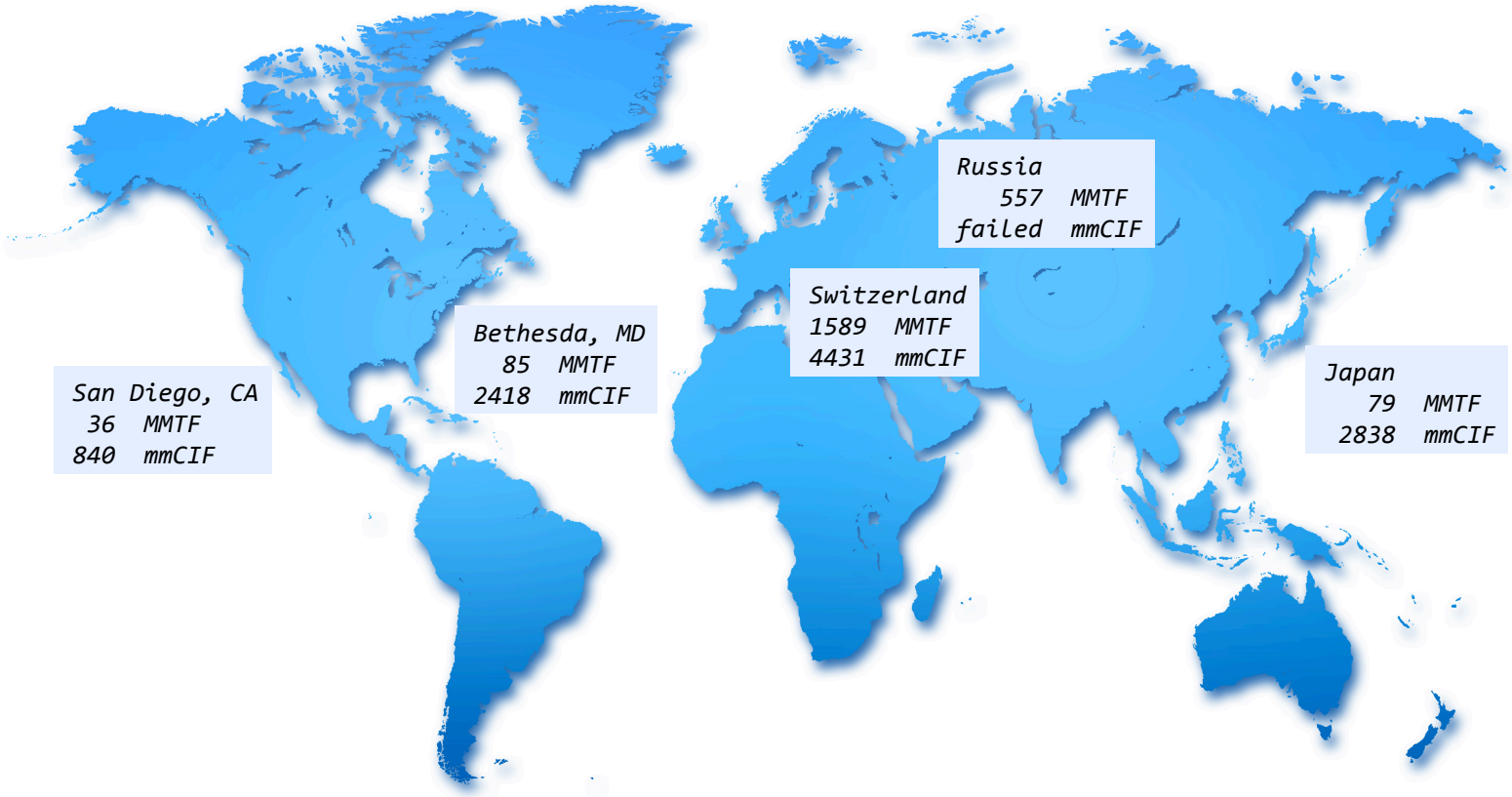
File Parsing Speed (PDB archive ~127,000 entries)



Comparison made with BioJava and mmtf-java API (single core)

Download + Parsing time MMTF vs. mmCIF

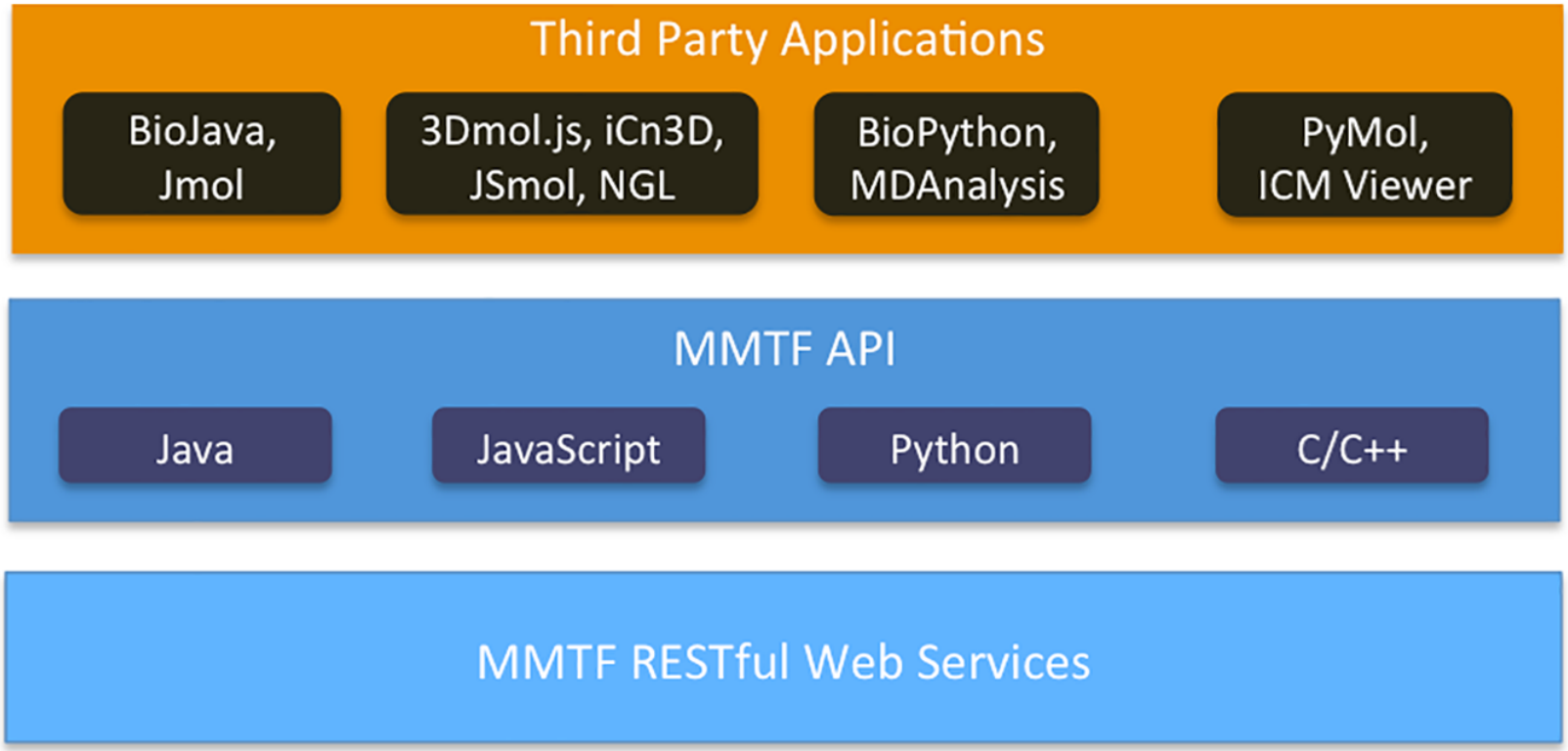
Time (seconds) to download* 100 large PDB structures from UCSD and parse with JavaScript decoder in Chrome browser



*Note: download times are highly variable and not representative

Community Engagement

- Open source specification
- Open source decoding libraries



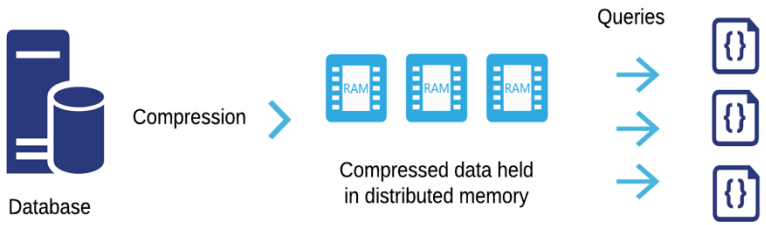
Summary

- **MacroMolecular Transmission Format (MMTF, mmtf.rcsb.org)**
  - Compressed, binary, efficient representation of 3D structures
    - Lossless representation (~4x compression)
    - Lossy, reduced representation (~37x compression)

- **Compressive Structural Bioinformatics**
  - Algorithms, application, and workflows using MMTF
    - 10 to 100+ fold speedup

Structure Visualization

Large Scale PDB Mining



Web-based molecular graphics for large complexes (2016)
Web 3D '16, 185-186, DOI: 10.1145/2945292.2945324

Resources

- **MMTF Website**
  - http://mmtf.rcsb.org
- **GitHub Repository**
  - https://github.com/sbl-sdsc/mmtf-spark
- **Publications**
  - Bradley AR, et al. (2017) MMTF—An efficient file format for the transmission, visualization, and analysis of macromolecular structures. PLOS Computational Biology 13(6): e1005575. https://doi.org/10.1371/journal.pcbi.1005575
  - Valasatava Y, et al. (2017) Towards an efficient compression of 3D coordinates of macromolecular structures. PLOS ONE 12(3): e0174846. https://doi.org/10.1371/journal.pone.0174846
  - Rose AS, et al. (2016) Web-based molecular graphics for large complexes. In Proceedings of the 21st International Conference on Web3D Technology (Web3D '16). ACM, New York, NY, USA, 185-186. https://doi.org/10.1145/2945292.2945324

Funding

This workshop was supported by the National Cancer Institute of the National Institutes of Health under Award Number U01CA198942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.